Mozout

Generated on 27/11/2026 08:09

KNOWLEDGE BASE · PUBLIC DOCUMENT

Using Claude Code with All in One AI: the variables, the model and common errors

All in One AI

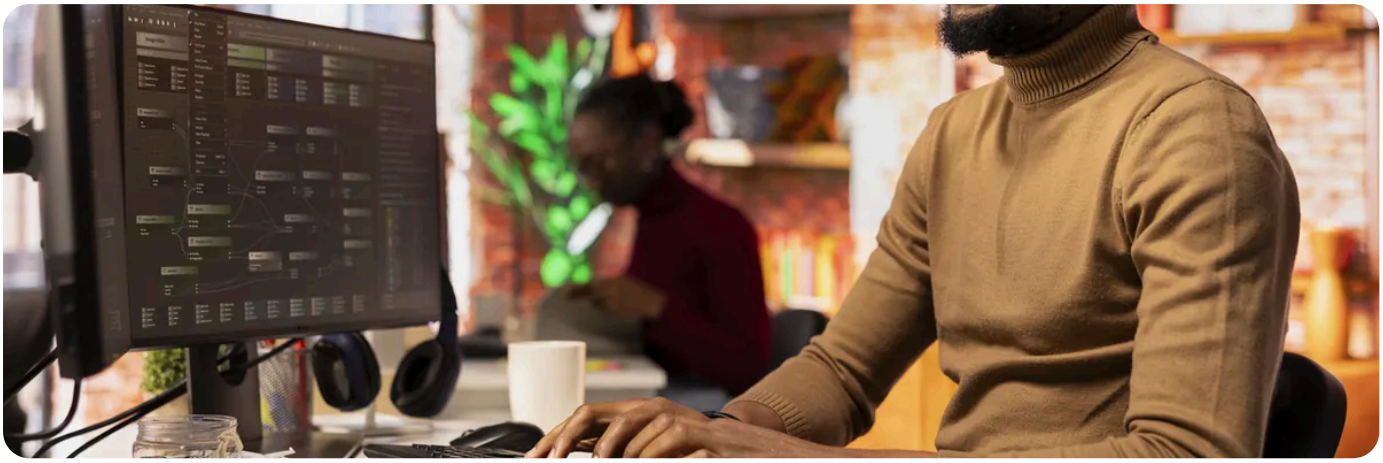
CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-8766	v2026.11.26	International · English

MozOut International
RLX Holdings (Pty) Ltd. · Group RLX Holdings
Tax no. (SARS) 943 907 6234 · Registration no. (CIPC) 2025/667953/07

https://www.my.mozout.com/knowledgebase/8766

Using Claude Code with All in One AI: the variables, the model and common errors

All in One AI Created on **November 26, 2026** 3 min read



Claude Code can work with All in One AI models instead of an account with the model maker. You pay with AI credits or with the Coding Plan, using the same MozOut API key. All it needs is the address, the key and the model, set as environment variables. The other tools are on the **Coding with AI** page.

Before you start

1. Create a key in the **API Center** with the `ai:read` and `ai:write` permissions. Use a `sandbox_` key to check the setup and a `prod_` key to work. The fields are explained in **creating an API key**.
2. Have AI credits or an active Coding Plan in **All in One AI**.
3. Keep the key in an environment variable, so it is not written in files: `export MOZOUT_API_KEY="prod_..."`. To have it in every new terminal, add the line to `~/.bashrc` or `~/.zshrc`. On Windows, use `setx MOZOUT_API_KEY "prod_..."` and open a new terminal.

The variables

Set them before you open Claude Code:

```
export ANTHROPIC_BASE_URL="https://api.mozout.com"
export ANTHROPIC_AUTH_TOKEN="$MOZOUT_API_KEY"
export ANTHROPIC_MODEL="moonshotai/kimi-k2.7-code"
export ANTHROPIC_DEFAULT_HAIKU_MODEL="deepseek-ai/deepseek-v4-flash"
export CLAUDE_CODE_MAX_CONTEXT_TOKENS=262144
export CLAUDE_CODE_MAX_OUTPUT_TOKENS=16000
claude
```

Variable	What it does
`ANTHROPIC_BASE_URL`	Our address, **without** `/v1`. Claude Code adds `/v1/messages` itself.
`ANTHROPIC_AUTH_TOKEN`	The key. You can also use `ANTHROPIC_API_KEY`: we accept both.
`ANTHROPIC_MODEL`	The main model. Change it during a session with `/model`, for example `/model zai-org/glm-5.2`.
`ANTHROPIC_DEFAULT_HAIKU_MODEL`	The model for background tasks, which should be fast and cheap. Older versions use `ANTHROPIC_SMALL_FAST_MODEL`.
`CLAUDE_CODE_MAX_CONTEXT_TOKENS`	The model's context, so Claude Code summarises the conversation at the right time.
`CLAUDE_CODE_MAX_OUTPUT_TOKENS`	The maximum per answer. Without it, Claude Code asks for 32,000 and we reserve credits for 32,000. A lower value needs less balance.

So you do not repeat the variables

Put them in `~/.claude/settings.json`, under `env`. Here the key is written out, because Claude Code copies the values exactly as they are. Keep this file on your own computer only.

```
{
"env": {
"ANTHROPIC_BASE_URL": "https://api.mozout.com",
"ANTHROPIC_AUTH_TOKEN": "prod_...",
"ANTHROPIC_MODEL": "moonshotai/kimi-k2.7-code",
"ANTHROPIC_DEFAULT_HAIKU_MODEL": "deepseek-ai/deepseek-v4-flash",
"CLAUDE_CODE_MAX_CONTEXT_TOKENS": "262144",
"CLAUDE_CODE_MAX_OUTPUT_TOKENS": "16000"
}
}
```

The right model

`moonshotai/kimi-k2.7-code` is the starting point for everyday work. If Claude Code asks for a name that is not in the catalogue, such as `claude-sonnet-...`, we use `moonshotai/kimi-k2.7-code` and charge its price. So always set `ANTHROPIC_MODEL`. The other recommended models are in **the right address and model in each tool**.

Common errors

Answer	When	What to do
`401`	The key is missing or wrong, or the terminal did not read the variable.	Check the key and open a new terminal after setting the variable.
`404 not_found`	The address has an extra `/v1`.	Use the address without `/v1`.
`402 insufficient_credits`	There are not enough credits for the answer's reservation.	Buy credits or lower `CLAUDE_CODE_MAX_OUTPUT_TOKENS`.
`429 plan_limit_reached`	The plan points have run out and the request cannot be paid with credits.	Wait for the reset, or turn on Use credits when the plan runs out.
`429 rate_limited`	You went over the key's requests per minute.	Wait a few seconds. Claude Code usually retries on its own.

In Claude Code, errors arrive in the Anthropic format: the `type` matches the HTTP status, for example `rate_limit_error` on a 429. The limits are in **API request limits**.

Check the setup with a `sandbox_` key.
Claude Code always gets "Resposta de teste (sandbox)." and nothing is spent. If that answer shows up, the address, the key and the model are right. Then switch to the `prod_` key: with the sandbox one, Claude Code cannot write code.

RECOMMENDED

Want to code with AI using our models? Visit the Coding with AI page.

See Coding with AI →

SEE ALSO

- **Cursor, Codex CLI, Cline and other coding tools: the right address and model in each** →
- **Coding Plan: which plan to choose, what it costs and how to subscribe** →

- The All in One AI MCP server: images, video and voice inside Claude Code, Codex and ChatGPT
- Testing All in One AI in the Sandbox at no cost: what each answer looks like
- Coding with AI in All in One AI

MozOut International

RLX Holdings (Pty) Ltd. · Group RLX Holdings

TAX NO. (SARS)	943 907 6234
REGISTRATION NO. (CIPC)	2025/667953/07
REGISTERED OFFICE	1111 Olive Street Saint Louis 63101, Missouri, United States
CONTACT	info@mozout.com · +27 78 823 3325 · www.mozout.com
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-INT-KB-8766 · v2026.11.26 · Generated on 27/11/2026 08:09